DOCUMENT RESUME

ED 159 224 TM 007 708

AUTHOR Mandeville, Garrett K.
TITLE An Evaluation of Title I Model C1: The Special Regression Model.
PUB DATE Mar 78
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Analysis of Covariance; Intermediate Grades; *Mathematical Models; Norm Referenced Tests; *Program Effectiveness; Program Evaluation; Reading Achievement; *Reliability; Research Design; *Simulation; Standard Error of Measurement; *Statistical Bias
IDENTIFIERS *Elementary Secondary Education Act Title I; *RMC Models

ABSTRACT

The RMC Research Corporation evaluation model C1--the special regression model (SRM)--was evaluated through a series of computer simulations and compared with an alternative model, the norm referenced model (NRM). Using local data and naticnal norm data to determine reasonable values for sample size and pretest posttest correlation parameters, the simulations used normal curve equivalent (NCE) scores with 200 repetitions for 10 sample sizes, three levels of correlation, and three ability levels. The results indicate that: (1) the standard errors of the estimates of project effectiveness are quite large in many cases and have a strong dependence upon the size of the comparison group; (2) as implemented through NCE scores, the procedure has a small bias; and (3) when project effectiveness estimates are obtained using the inappropriate norm referenced model, the bias may be large enough to suggest that the results are educationally significant even in cases where there is no real effect. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159224

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Garrett K. Mandeville

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

AN EVALUATION OF TITLE I MODEL C1

THE SPECIAL REGRESSION MODEL

Garrett K. Mandeville

University of South Carolina

Paper presented at the annual meeting of the
American Educational Research Association
Toronto, Canada, March 27-31, 1978

Printed in the United States

TM002708

Session 9.13

Abstract

In this study standard errors of estimates of project effectiveness are reported for a Title I evaluation Model, which because it requires that the identification of project participants be based on low performance on a pretest, should have widespread appeal. Using local data and national normative information to determine reasonable values for the parameters (sample sizes and pretest-posttest intercorrelations), a simulation study was conducted. The results indicate that (1) the standard errors of the estimates of project effectiveness are quite large in many cases and have a strong dependence upon the size of the comparison group, (2) as implemented through NCE scores the procedure has a small bias and (3) when project effectiveness estimates are obtained using the inappropriate Norm Referenced Model, the bias may be large enough to suggest that the results are educationally significant.

Introduction

It is common to select participants for a special instructional innovation based on a measured deficiency they possess. This is particularly true of programs supported by federal funds such as ESEA Title I but need not be limited to such situations. For example, without federal support, a local district might determine to set aside special funds to aid students with an educational deficiency. In either case, sound logic and financial considerations suggest that the educational intervention employed in an attempt to remediate the selected students should be evaluated.

Assuming that all students in the educational agency, e.g., school, which are identified as deficient are to be remediated, standard experimental designs based on randomization are impossible to employ. It is for these situations that the RMC Research Corporation (1976), under contract to the Office of Planning, Budgeting and Evaluation of the Office of Education, suggested the use of a regression model to evaluate program effectiveness. Although the evaluation model, termed Model C1 - The Special Regression Model (SRM), has "face validity", only very general guidelines were provided for researchers considering its use. This paper presents a systematic study of the characteristics of the model and compares it to an alternative model, the Norm Referenced Model (NRM) which, although inappropriate, may appear attractive to evaluators.

Perspectives and Theoretical Framework

The SRM is prescribed for those situations in which project participants are identified by scoring below some cut-off score on a test, a weighted combination of test scores, or a composite performance index which, for example, might include teacher ratings. This, of course, requires that all potential students are measured on whatever variables are relevant to the selection process prior to actual selection. It is also necessary that selection for project participation be strictly a function of the score or composite score: all students falling below the criterion must be included in the special program while none of those with scores above the cut-off may

be included. Hereafter the former group will often be referred to as the treatment or treated group and the latter as the comparison group.

The special program is then applied to the project participants whereas the remainder of the students receive the normal instructional program. When the program under evaluation has concluded, all students in both groups are posttested. Although the posttest would generally be the same as or a parallel form of the pretest, this is not necessary. It should, of course, be content valid with respect to the objectives common to the two instructional programs. To estimate project effectiveness neither the pretest nor the posttest are required to be standardized tests. The three models presented by RMC, however, require that project effectiveness estimates be provided in a metric which requires normative data. That is to say that locally developed tests would suffice to provide data on project effectiveness but would not, unless a standardized test were used in addition to the locally produced test, provide project effectiveness estimates required by Title I.

The first step in obtaining the estimate of project effectiveness is to regress the posttest onto the pretest for the comparison group and extrapolate this regression line into the region of pretest scores obtained by the treated group. The assumption is that the regression line provides an estimate of the trend in performance which would have been obtained by project participants if they had not been in the project, i.e., if they had received the regular instructional program. From another point of view, we are assuming that if the project were ineffective a regression line based on the treatment group would, within sampling error, coincide with the one based on the comparison group. If the project has been effective, however, and no interaction of treatment and pretest occurred, the regression line for the treated group should be approximately parallel to the comparison group regression line but shifted above it. The estimate of project effectiveness is calculated as the difference on the posttest between the two regression lines at the mean pretest score for the treated group. (Note that this is merely the difference between the "prediction" from the comparison group regression line and the mean

posttest for the treated group.) As a check of the consistency of the effect, it is also suggested that it be estimated at the cut-off score used to differentiate treatment and comparison group students.

Although it is to be expected that SRM estimates are unbiased, some factors which should affect the standard error of the estimates are the size of the comparison group, the pretest-posttest intercorrelation, and the extent of the extrapolation of the regression line. On the first two points RMC suggests "ideally about 50 treatment and 100 comparison pupils" and "a high correlation can be reasonably expected between pre- and posttest scores" (1976, p. 29). It might be expected that the size of the comparison group is the more critical of the two sample sizes because of its relationship to the stability of the regression line. The size of the pretest-posttest intercorrelation is also related to the stability of the regression. Of more direct relevance, however, is the fact that if this correlation is not of moderate size the procedure is not feasible. Since the standard errors of regression estimates are directly related to the distance between the mean of the predictor for the sample on which the equation is based and the value of the predictor, estimates for more extreme treatment groups should be less stable.

The Norm Referenced Model is another of the three evaluation models proposed by RMC. Proper application of the NRM requires that project participants not be selected on the basis of pretest performance or, if so, that another pretest be administered for purposes of providing data for the estimation procedure. The NRM uses the national norming sample as the comparison group and, therefore, is most appropriate when the local population is similar to the norm group. The assumption of the NRM is that, if the project is ineffective, the treatment group should maintain its same position relative to the national norms from pretest to posttest. Thus, the mean score of the treated group would be at the same national percentile on both the pretest and posttest. Assuming that the project is ineffective, the prediction of the mean posttest score for the treated group is the posttest score which has the same percentile rank as their pretest mean. The difference between the observed mean

posttest and the prediction provides the estimate of project effectiveness.

The requirement of the NRM that students not be selected for participation based on the pretest is to guard against the regression effect which would tend to inflate estimates of project effectiveness. This bias, of course, is inversely related to the size of the pretest-posttest intercorrelation but is directly related to the extent to which the group is "extreme", i.e., does not compare favorably with the norm group.

Methods and Procedures

In this paper estimates of project effectiveness were obtained in situations where students were selected for participation based on low pretest scores, i.e., for situations in which the SRM is appropriate. In addition to providing estimates for the SRM, estimates were also obtained for the NRM. These latter estimates were produced to provide information on the extent to which project effectiveness is overestimated when the NRM is applied inappropriately. The procedures employed in this investigation will be presented in two parts: (1) the development of programming which was used in the simulation process and (2) the selection of the parameters for the simulation. Although these two phases of the study are interrelated, the rudiments of the programming task will be discussed initially.

Characteristics of the Simulation Program

For simplicity it was assumed that the distribution of pretest - posttest scores would follow a bivariate normal distribution. This is not an unreasonable assumption since many evaluations are based on standardized tests which provide an expanded scale score or standard score metric and these are constructed to be normally distributed within each grade level. Also, the appropriateness of the two models considered depends upon there being no severe ceiling or floor effects. Finally, from a practical standpoint, bivariate normal data are rather easy to simulate. Thus, the computer program was written

to allow input of the five parameters required to describe a bivariate normal distribution to provide the pretest - posttest score pairs for some local school population.

The next issue dealt with was the metric of these initial scores. The discussion of the three models in the RMC publication, although vague, suggests that standard scores will be used during the preliminary calculations, e.g., in the computation of means and regression coefficients, and that these will be converted to the preferred metric, normal curve equivalents (NCE's), in the final estimation stage. (Note: NCE's are normally distributed with a mean of 50 and a standard deviation of 21.06 and provide interval level measurement.)

It is clear that the RMC recommendations were made in order to minimize the computational effort of local school personnel. There is an implicit assumption, however, that the conversion from the initial metric to NCE's is a linear transformation. It is unlikely that this assumption is valid in general; it must depend upon the type of "initial" score used and the test publisher's development of that score. In order to write a general program which did not depend upon the nature of the transformation from initial scores to NCE's, a decision was made to convert these scores to NCE's and carry out computations in this metric. Conversion to NCE's, however, required the intermediate step of converting the scale scores generated to percentile ranks; these were then converted to NCE's using an inverse normal transformation.

Because of the availability of local data, expanded scale scores on the Comprehensive Test of Basic Skills, Form S (CTBS/S, 1974) were considered as the base data and the total reading score was selected for this study. Therefore scale score to percentile rank normative information for the years 4.7 (pretest) and 5.7 (posttest) was read into the computer to facilitate the direct conversion of the scale scores generated to NCE's. A definite advantage in beginning with scale scores is that this is a fairly common metric in which summary data are available, thus, aiding in the selection of reasonable input parameters.

Given a set of bivariate data in an NCE metric, the score pairs were then sorted on the basis of pretest scores. It was assumed that available resources allowed for a fixed number of subjects to be selected for project participation and, in keeping with the SRM, those students with the lowest pretest scores were selected.

The posttest on pretest regression was then obtained for those students not selected, i.e., the comparison group. The predicted posttest score from this regression line was then obtained at the pretest mean for the treatment group; when this was subtracted from the observed posttest mean for the treated group, the result was the estimate of project effectiveness in NCE's. The whole process was repeated 200 times for each set of parameters.

Several statistics descriptive of these estimates were obtained over the 200 simulations. These included the average of the estimates, the standard error and square root of the mean square error and indices of skewness and kurtosis of the distribution.

All data sets simulated included zero treatment effects, i.e., represent situations in which the project was not effective. There was no loss of information associated with this decision because inclusion of non-null treatment effects would merely have translated the average estimate by that amount and, of course, other characteristics of the distribution would remain unchanged.

The estimates of project effectiveness for the NRM were also obtained simply by taking the difference between the posttest and pretest means for the treated group. With the exception of the statistics dealing with the form of these distribuions, the statistics calculated for SRM estimates were also obtained for those based on the NRM.

The programming was accomplished in FORTRAN IV with IMSL (1977) subroutines GGNRM (and GGNRM1) and MDNRIS used to provide samples from a bivariate normal distribution and the inverse normal transformation, respectively. The latter routine was needed to convert percentile rank scores to NCE's.

The estimation procedures described above are at variance with those presented in the RMC manual. RMC does not specify what type of

scores should be used except for an occasional reference to standard scores. The implication in the RMC document is that it makes no difference when the conversion to NCE's is made and in order to ease the computational effort at the local level, the conversion should be the last step in the process.

Parameter Selection

As noted above, the pretest and posttest scores simulated in this investigation were patterned after the Total Reading score distribution of the CTBS/S. Because data on a large sample of students (N=1895) from a local school district were available, and the tests were administered in the spring of the 4th and 5th grades, pretest and posttest scores were simulated to correspond to scale scores on the test at the 4.7 and 5.7 norms, respectively.

In the selection of parameters describing the bivariate score distributions to be generated, attempts were made to simulate both national and local populations. To this end, two initial parameter sets were selected, one based on descriptive statistics from the CTBS Technical Bulletin (1974) and the other based on the same information for the local sample. For the national sample the mean values were taken to be 420 and 460 (Tables 30 and 31 for levels 1 and 2) for the pretest and posttest, respectively; the standard deviations used were 79 and 85 and the pretest-posttest intercorrelation was .72 (Table 53, level 1 and 2 intercorrelation). The corresponding statistics based on the local data were means of 360 and 420, standard deviations of 77 and 88 and a correlation of .87.

Since we expect the stability of the SRM and the bias of the NRM estimation procedures to be a function of the degree of difference between the local population and the national one, it was determined that a reasonable parameter to systematically vary would be average performance levels on the pretest and posttest. Therefore, scale scores corresponding to P_{10}, P_{30} and P_{50} were obtained for each of the two grades (4.7 and 5.7) from the Technical Bulletin. Thus, the input means for each of these five parameter sets were consistent with an assumption of the NRM. In selection of the standard

deviations to correspond to these five performance levels, the local data were run for each of the 16 schools in the sample. It might be expected that for schools lower than the national norms, the variation would be reduced due to floor effects but this was not found to be the case. The school standard deviations were quite consistent, ranging between 70 and 80 for the pretest and 75 and 95 on the posttest. For this reason, integer values nearest to the median standard deviation were selected to be used for all simulations and these values were 77 and 86 (which, by the way, are virtually identical to the national values of 79 and 85).

The earlier comparison of national and local data indicated a higher pretest-posttest intercorrelation in the local data than was indicated in the national sample. Therefore, in keeping with the interest in simulating both local and national data, the median correlation for the 16 schools of .86 was used as one level of this parameter and the value of .72 from the Technical Bulletin was used as a second. Since it is clear that both procedures will suffer if the intercorrelation is decreased, a third value of .60 was also employed. Thus, the major portion of the simulations were based on three levels of pretest-posttest location parameters and three levels of pretest-posttest intercorrelations, or 9 parameter sets in all.

We also know that the absolute and relative sizes of the treated and comparison groups are important to the results. The RMC recommendation was for 50 treated and 100 comparison students. In this research situations in which the combined sizes of the treated and comparison samples ranged from 60 to 120 were simulated. This is in keeping with typical elementary school grade sizes, at least as reflected in the local data where 12 of the 16 schools had 125 or fewer students. Although these figures are underestimates of actual 4th and 5th grade enrollments, due to mortality in the matching of the longitudinal data for the two years, the sample sizes were determined to be reasonable because of the anticipated degredation of the procedures for small sample sizes. Computer time was also a consideration here. Thus schools with populations of 60, 80, 100 and 120 students were simulated.

For schools with 60 students, it was felt that one class of 20 students might be selected for the project or, if funds permitted and/or the need were great enough, 30 students might participate. For schools with 90 students, the same sizes were used for the treated samples, leaving 20 more children for each comparison group. For schools with 100 students, treated sample sizes of 20, 30 and 40 were used and, to introduce a little variation, treatment N's of 20, 40 and 60 were used for schools with 120 students.

Results

The results for the ten computer simulations patterned after the national norm group are presented in Table 1. The quantities denoted in the table are defined below and will be used in the remainder of the paper.

N_{TRT} is the sample size for the treated group

N_{NON} is the sample size for the comparison group

N_{ALL} is the combined sample size

$\bar{E}_S$ is the average of the 200 SRM estimates

S_{E_S} is the standard deviation of the 200 SRM estimates

SK_{E_S} is the skewness index of the 200 SRM estimates and is calculated as $\frac{1}{200}\sum(E_S - \bar{E}_S)^3 / S_{E_S}^3$

KR_{E_S} is the kurtosis of the 200 SRM estimates and is calculated as $\frac{1}{200}\sum(E_S - \bar{E}_S)^4 / S_{E_S}^4 - 3$

$\bar{E}_N$ is the average of the 200 NRM estimates

S_{E_N} is the standard error of the 200 NRM estimates

$SMSE_N$ is the square root of the mean square error of the 200 NRM estimates which, since the true effect size is zero, is calculated as $\frac{1}{200}\sum E_N^2$

Looking first at the form of the distributions of SRM estimates, we observe the values of skewness and kurtosis are tolerably close to zero, the values appropriate for the normal distribution. According to McNemar (1962), the standard error of the skewness index from a normal distribution is $\sqrt{6/N}$ which, for N=200, is .173. Using a two standard error criterion of .346, we note that all ten values of SK_{E_S} are, in absolute value, acceptably close to zero. Again citing McNemar, the appropriate comparison value for the index of kurtosis is twice the

value for skewness, or .692. Only one value of KR_{E_S} (N_{TRT} = 20, N_{NON} =100) exceeds this value. Therefore it appears as though a normal model for the E_S distribution is reasonable, at least for this set of parameters.

With respect to bias in E_S all ten values of $\bar{E}_S$ are within two standard errors of zero. The largest (absolute) z value is -1.77 ($-.63x\sqrt{200}/5.02$) for N_{TRT} = 30 and N_{NON} = 70. Because of the minimal bias in the E_S estimator, the mean square errors for E_S have not been presented.

The standard error of the SRM estimates are rather clearly related to the size of the comparison group. The correlation between S_{E_S} and N_{NON} for the ten data points is -.86 indicating a reduction in the standard error as the comparison sample size increases, an expected result. The sizes of these standard errors are disturbingly large, however; in only one case was S_{E_S} smaller than 5. In the NCE metric, with a standard deviation of 21.06, an S_{E_S} of 7 indicates, assuming normality, that roughly one-third (32%) of the estimates will be about 1/3 or more standard deviations away from zero, even if the treatment is no more effective than the instruction provided for the comparison group.

In an earlier publication (1975, p. 53) the authors of the RMC publication suggested that a treatment effect might be considered "educationally significant" if it were 1/3 of a standard deviation or larger. The implication was that this was in addition to statistical significance. For small sample applications of the SRM we see that an evaluator is quite likely to obtain estimates which appear by the above standard to be educationally significant but which are well within sampling error of zero.

Because of the high correlation between N_{NON} and S_{E_S}, a regression equation for predicting the standard error from the comparison group sample size was obtained. On intuitive grounds, the square root of N_{NON} was used rather than the value itself and two other runs were made (50, 100 and 30, 120 were the sample sizes) in an attempt to more adequately describe the relationship. The new runs with a total sample size of 150 produced standard errors near 4.3 and

the correlation between $\sqrt{N_{NON}}$ and S_{E_S} was -.93 (it was -.90 using N_{NON}).

The regression equation was

$$S'_{E_S} = 9.98 - .535(\sqrt{N_{NON}})$$

The size of the correlation indicates that the above equation could be used for situations similar to the one simulated.

Looking at the estimates from the NRM we note the positive bias due to the regression effect. In the NCE metric, values of $\bar{E}_N$ range between .25 and .45 standard deviations for the ten sample sizes simulated. In addition to the bias in the NRM estimates, they are also much more consistent with standard errors in the 2 to 3 point range. Unfortunately, in every case a positive E_N is almost a certainty.

That the size of the bias is inversely related to the proportion of the lower tail of the pretest distribution which is treated is demonstrated by the fact that the correlation between $N_{FRAC} = (N_{TRT}/N_{ALL})$ and $\bar{E}_N$ is -.98. A correlation of this magnitude suggests the possibility of a linear equation to predict the bias as a function of N_{FRAC}. The resulting equation is

$$\bar{E}'_N = 11.60 - 12.85(N_{FRAC})$$

Although extrapolation beyond the sample sizes used here would be risky, it would be reasonable for an evaluator to use the above equation to obtain an estimate of the expected bias if the NRM had been applied to data reasonably similar to the situation simulated.

For example, suppose a special reading project for 25 5th graders out of a population of 100 produced an NRM estimate of 10 NCE's. Assuming the population is similar to the national norming sample (required for application of the NRM anyway) and that the CTBS/S had been used, we could predict the value of $\bar{E}_N$ to be

$$\bar{E}'_N = 11.60 - 12.85(.25) = 8.4$$

From the tabled values of S_{E_N} for (20,80) and (30,70) which are 3.24 and 2.83, respectively, we might expect a standard error of about 3 NCE points. Thus the observed value of 10 is less than .5 standard errors above expectation which, by most scientific standards, is insufficient evidence that the project was effective.

The results for the simulations of local data are presented in Table 2. Once again the E_S distributions tend to be normal with respect to skewness and kurtosis. A negative bias in the estimator is indicated for two of the ten sample sizes, however. The (20,80) sample with $z = -1.78$ is also approaching significance. The reason for this is artifactual, however, and can be explained as follows.

Scale score to percentile rank conversions in the computer program were made using norm tables prepared by McGraw-Hill and, according to the tables, the 1st percentile at grade 4.7 is a scale score of 255. (This was not verified in the local data since scores as low as 150 were observed for this grade level.) Therefore, approximately 8% of the "local" pretest scores generated would fall below this value and create a "floor" effect. The corresponding percentage for the posttest is about 5%. Naturally, the floor effects will occur only for the subjects assigned to the treated sample and the smaller the proportion of treated students, the larger the percentage of their scores which will be at the floor. Since the regression model is clearly inappropriate in this situation, the procedure evidently fails.

We also note that the larger correlation has caused the standard error of E_S to decrease as anticipated. The bias in E_N, on the other hand, has not decreased as would be expected when the regression effect is reduced. This, however, is probably because the "null case" assumption of the NRM has been violated. Student performance from pretest to posttest has increased since the percentile ranks associated with the scale score means are approximately 22 and 32 respectively. The NRM estimates, therefore, reflect a true treatment effect in addition to the contribution of the regression effect.

Table 3 contains partial results for the 30 runs made using pretest and posttest means corresponding to the 50th percentiles in the norms. In contrast to the initial simulations representing national data we note a positive bias in E_S, particularly for small treatment sample sizes (20 or 30) and low pretest-posttest intercorrelation (.60). In addition to the three values of $\bar{E}_S$ which are significantly different from zero, two others are positive and

quite large. All $\bar{E}_S$ are positive for the correlation of .60, an unlikely event if the procedure is unbiased. For the larger pretest-posttest intercorrelations, however, there are no trends in the data to suggest that E_S is biased. Standard errors demonstrate the expected relationship to the intercorrelation but are not reduced much below 4 NCE points even under the most ideal conditions simulated.

The bias in E_N is even more strongly affected by ρ, being 10 or more NCE points on the average when the treated group is most extreme and $\rho = .60$. On the other hand, when the pretest and posttest are highly related ($\rho = .86$), the regression effect and, therefore the bias, is quite small.

The results presented in Table 4 do not speak well of the behavior of the SRM if the population sampled lies below the norm and pretest to posttest progress is in accordance with the norms. The estimates are severly positively biased except when $\rho = .60$ where the results are inconsistent.

To explain this result we must reconsider the initial reasoning behind the selection of input parameters, in particular, the means. They were chosen in accordance with the NRM assumption that the population would maintain the same national percentile from pretest to posttest. The regression effect, on the other hand, guarantees that this will not be true for an extreme group such as one which is designated to be at the 30th percentile on the pretest. In this case "normal progress" depends on the pretest-posttest intercorrelation and can be estimated to be posttest mean scale scores of approximately 431, 426 and 419 for the correlations used in the simulations, rather than the constant of 415. Thus it appears as though the fact that the norms tables used in the program are not appropriate for the data simulated, a regression effect of sorts operates. Possibly the reason it is not as strong for $\rho = .60$, where logic would suggest it should be most extensive, can be explained by the lack of stability in the regression line.

The most difficult aspect of these results is in determining whether they signify inappropriate selection of input parameters in

relation to the norms tables used, a fault of the estimation method, or both. It does seem possible, even likely, that populations similar to those simulated do exist; if so, it is clear that SRM estimates based on the normative information provided by the test publishers, are positively biased. Fortunately, however, when compared to their standard errors, the bias in E_S is small.

Standard errors of SRM estimates in Table 4 are very similar to those in Table 3 and require no comment. The bias in NRM estimates for these populations is generally slightly larger than for those exhibited in Table 3.

As might be expected, Table 5 also presents a rather negative picture of the behavior of E_S when the populations sampled are more deviant from the norms. Here, however, a negative bias is found when the treatment sample size is small and there are 80 or more observations in the comparison group. These are precisely the situations for which the floor effects, described in the discussion of Table 2, are operating. Estimating that approximately 18% of the pretest scores would be at the floor (scale score of 255), it is clear that the regression model is not appropriate. No other aspects of Table 5 are noteworthy.

Summary

The results indicated that when the simulation parameters agreed precisely with the normative data, the behavior of SRM estimates was exemplary. They were unbiased and, at least with respect to 3rd and 4th moments, appeared to be normally distributed. Standard errors, however, were larger than one would like, being as large as .5 standard deviations in the NCE metric for small samples. Thus, evaluators must demonstrate considerable caution in attributing even fairly large estimates of treatment effects to project effectiveness.

The expected dependence of the standard errors on the size of the comparison group was documented and a regression equation was presented. The consistency of the sizes of standard errors across the various simulations was noted. Although not presented, linear regression equations for predicting S_{E_S} from N_{NON} were obtained for

all sets of 10 computer simulations. Pearson r's were typically in the -.90's and, for a given pretest-posttest intercorrelation, the coefficients in the regression equations were quite stable. After rounding, the following equations fit the data reasonably well:

$$\rho = .60 \quad S'_{E_S} = 11 - .06N_{NON}$$
$$\rho = .72 \quad S'_{E_S} = 9 - .05N_{NON}$$
$$\rho = .86 \quad S'_{E_S} = 6 - .03N_{NON}$$

Because these represent "average" regression equations over rather varying conditions, and they are in the NCE metric, it is suggested that evaluators may use them to obtain estimates of standard errors of E_S for designs they might contemplate.

Although anyone with an introductory statistics course is familiar with the regression effect, its magnitude as quantified in this study may be shocking to evaluators who have voluntarily or involuntarily had to employ the NRM in this inappropriate situation.

For the simulations based on local data, the results appeared to "track" rather well except for those sample sizes for which the floor effect rendered the procedure invalid. This is curious since the local parameters are similar to those in Table 4 (pretest and posttest means maintained at the 30th national percentile) for $\rho = .86$ where a positive bias in E_S was identified. These and the remaining results indicated a problem with the procedure; it was suggested that the likely cause of this problem was the dependence upon the fit of the local population to the national norming sample.

Possibly the problem is artificial; it surely would be reduced or eliminated completely if the conversion through the norms to NCE's were not required. Statistically speaking the procedure is surely sound. When the psychometric niceties of NCE's enter the picture, however, it appears as though problems arise.

The argument that school populations such as those simulated do not exist was invalidated by reference to the local data. Although the results for the local data appeared reasonable, the fact that they are at variance with those for a similar set of parameters caused some loss of faith in the procedure.

Conclusions

The writer began by assuming that the SRM was an unbiased estimation method and expected to demonstrate this fact. The expected contribution of this paper was to provide an indication of the magnitude of the standard errors of these estimates in order to aid evaluators in judging their "statistical" significance. As an aside, the bias of the NRM was to be demonstrated and quantified.

The results do not support the expectation of unbiasedness of SRM estimates. The bias is not large, however, particularly when compared with the bias of NRM estimates. Therefore, in spite of the negative findings reported here, the SRM is clearly superior to the NRM in situations where project participants are selected on the basis of a pretest.

The regression equations for predicting standard errors from comparison group sample sizes should prove useful to evaluators. The consistency of these results across Tables 3 through 5 indicate that these equations should provide at least the order of magnitude of standard errors for sample sizes and pretest-posttest intercorrelations in the ranges simulated. Since the NCE metric was used, they should not be dependent upon the test.

The most troublesome aspect of the results is the demonstrated bias, albeit a small one, in the SRM estimates. It remains for statisticians and/or psychometricians to look more carefully at the procedure as implemented through NCE's, to determine the causes of this bias and whether it should concern the practitioner.

Table 1

Descriptive Statistics for SRM and NRM Estimates
in NCE's for 200 Simulations Representing
National Data - N(420, 460, 79, 85, .72)

N_{TRT}	N_{NON}	N_{ALL}	Special Regression Model (SRM)				Norm Reference Model (NRM)		
			$\bar{E}_S$	S_{E_S}	SK_{E_S}	KR_{E_S}	$\bar{E}_N$	S_{E_N}	$SMSE_N$
20	40	60	.51	6.65	.17	-.56	7.01	3.37	7.78
30	30	60	- .83	7.52	-.06	-.23	5.40	2.58	5.99
20	60	80	.29	6.10	.14	-.18	8.83	3.43	9.48
30	50	80	- .13	5.73	.05	.04	6.49	2.71	7.03
20	80	100	.25	5.66	-.31	.04	9.24	3.24	9.79
30	70	100	- .63	5.02	.23	-.25	7.35	2.83	7.88
40	60	100	.42	5.68	-.08	.03	6.54	2.43	6.98
20	100	120	- .34	5.00	-.20	1.02*	9.57	3.35	10.14
40	80	120	.09	4.93	.11	-.07	7.06	2.55	7.51
60	60	120	.03	5.63	-.02	-.27	5.42	1.82	5.72

*Significantly different from zero at the 5% level

Table 2

Descriptive Statistics for SRM and NRM Estimates
in NCE's for 200 Simulations Representing
Local Data - N(360, 420, 77, 88, .87)

N_{TRT}	N_{NON}	N_{ALL}	Special Regression Model (SRM)				Norm Referenced Model (NR		
			$\bar{E}_S$	S_{E_S}	SK_{E_S}	KR_{E_S}	$\bar{E}_N$	S_{E_N}	$SMSE_N$
20	40	60	.32	4.94	.29	-.45	8.39	2.31	8.70
30	30	60	- .46	5.45	.08	-.29	8.07	1.88	8.28
20	60	80	- .17	4.35	.05	-.05	8.98	2.35	9.29
30	50	80	- .10	4.29	.20	.17	8.27	1.95	8.49
20	80	100	- .50	3.96	-.28	-.01	8.73	2.17	9.00
30	70	100	- .55*	3.63	.19	-.26	8.41	1.97	8.63
40	60	100	.21	4.22	.00	-.09	8.38	1.63	8.54
20	100	120	-1.06*	3.57	-.17	.61	8.66	2.37	8.98
40	80	120	- .04	3.49	.15	.30	8.42	1.72	8.59
60	60	120	.06	4.04	-.05	-.21	8.10	1.30	8.20

*Significantly different from zero at the 5% level.

Table 3

Descriptive Statistics for SRM and NRM Estimates in NCE's for 200 Simulations Representing Data at the 50th Percentile N(422, 459, 77, 86, ρ_i) for ρ_i =.60, .72 and .86

N_{TRT}	N_{NON}	N_{ALL}	$\bar{E}_S$			S_{E_S}			$\bar{E}_N$		
			.60	.72	.86	.60	.72	.86	.60	.72	.8
20	40	60	1.32*	- .24	.28	8.56	7.34	4.87	8.25	5.65	2.4
30	30	60	1.05	- .43	- .90*	10.01	8.28	5.84	6.29	3.85	1.1
20	60	80	.30	1.18*	- .25	6.83	6.39	4.43	10.04	6.80	2.6
30	50	80	.82	.37	- .45	7.66	6.35	4.64	7.97	5.32	2.3
20	80	100	.27	.06	.13	6.27	4.72	3.79	11.12	7.36	3.4
30	70	100	.99*	.03	- .04	6.13	5.32	4.18	9.30	6.13	2.5
40	60	100	.14	.34	.36	6.28	5.56	4.62	7.67	5.06	2.4
20	100	120	.96*	.53	.44	5.57	4.97	3.65	11.70	8.13	3.1
40	80	120	.43	- .08	- .08	5.84	4.89	3.59	8.58	5.58	2.3
60	60	120	.15	.41	- .50	7.51	5.56	4.43	6.08	4.12	1.8

*Significantly different from zero at the 5% level.

Table 4

Descriptive Statistics for SRM and NRM Estimates in NCE's for 200 Simulations Representing Data at the 30th Percentile N (380, 415, 77, 86, ρ_A) for ρ_A =.60, .72 and .86

N_{TRT}	N_{NON}	N_{ALL}	$\bar{E}_S$			$S_{\bar{E}_S}$			$\bar{E}_N$		
			.60	.72	.86	.60	.72	.86	.60	.72	.86
20	40	60	.02	.64	.69*	8.45	7.37	4.82	8.84	6.42	3.08
30	30	60	-.10	.18	.85*	10.10	8.30	5.63	6.75	4.53	2.48
20	60	80	.56	.92*	.69*	6.67	5.79	3.94	10.35	7.62	3.65
30	50	80	.75	1.66*	.85*	7.02	5.95	4.42	8.68	6.22	2.84
20	80	100	.58	1.15*	.99*	6.18	5.01	4.02	11.81	8.14	4.05
30	70	100	1.22*	1.16*	.98*	6.26	5.22	3.78	9.8[illegible]	6.91	3.35
40	60	100	1.05*	1.29*	.82*	6.73	5.71	4.16	8.33	5.84	2.75
20	100	120	.56	1.53*	1.84*	5.37	4.53	3.22	12.46	8.89	4.81
40	80	120	1.14*	1.03*	.73*	5.62	4.51	3.75	9.33	6.37	3.23
60	60	120	.87	.11	.55	6.62	5.63	4.16	6.77	4.50	2.35

*Significantly different from zero at the 5% level.

Table 5

Descriptive Statistics for SRM and NRM Estimates in NCE's for 200 Simulations Representing Data at the 10th Percentile N(322, 352, 77, 86, ρ_i) for ρ_i =.60, .72 and .86

N_{TRT}	N_{NON}	N_{ALL}	$\bar{E}_S$			S_{E_S}			$\bar{E}_N$		
			.60	.72	.86	.60	.72	.86	.60	.72	.86
20	40	60	.53	.41	1.07*	7.37	5.71	4.07	8.65	6.29	3.57
30	30	60	2.06*	1.07	1.11*	8.71	8.27	6.11	6.94	4.99	2.45
20	60	80	- .78*	.14	.56*	5.38	4.86	3.63	9.06	6.62	3.57
30	50	80	.44	1.31*	1.68*	6.39	5.32	3.74	8.19	6.11	3.40
20	80	100	-1.72*	-1.30*	- .96*	4.18	3.84	2.48	8.88	6.27	3.05
30	70	100	- .01	.12	1.17*	4.99	4.49	3.40	9.20	6.47	3.72
40	60	100	.58	.83*	1.30*	5.38	4.82	4.01	8.01	5.83	3.25
20	100	120	-2.43*	-1.98*	-1.58*	4.18	2.98	2.68	8.43	5.73	2.65
40	80	120	.23	.68*	1.49*	4.94	3.98	3.25	8.89	6.38	3.65
60	60	120	1.17*	1.82*	1.43*	5.99	5.48	4.04	6.95	5.03	2.61

*Significantly different from zero at the 5% level.

REFERENCES

Comprehensive Tests of Basic Skills, Form S, Levels 1 and 2, Monterey, California: CTB/McGraw-Hill, 1974

Comprehensize Tests of Basic Skill, Form S All Levels, Technical Bulletin, Monterey, California: CTB/McGraw-Hill, 1974

Horst, D.P., Tallmadge, G.K. and Wood, C.T. A Practical Guide to Measuring Project Impact on Student Achievement, Washington: U.S. Government Printing Office, 1975

McNemar, Q. Psychological Statistics, New York: Wiley, 1962

Reference Manual for IMSL Library 1, Edition 6, Houston, Texas: International Mathematics and Statistics Libraries, 1977

Tallmadge, G.K. and Wood, C.T. USER'S GUIDE ESEA Title I Evaluation and Reporting System, Mountain View, California: RMC Research Corporation (for HEW), 1976

The writer is indebted to the administrative staff of Richland School District #1, Columbia, South Carolina for allowing access to data used in this report.